EE5110/EE6110: Special Topics in Automation and Control
Topic: State-variable Estimation

Arthur TAY

ECE, NUS

Outline

Introduction

- Sensors usually do not provide perfect and complete data about a system.
- First, they generally do not provide all the information we would like to know: either a device cannot be devised to generate a measurement of a desired variable or the cost (volume, weight, monetary, etc.) of including such a measurement is prohibitive.
- In other situations, a number of different devices yield functionally related signals, and one must then ask how to generate a best estimate of the variables of interest based on partially redundant data. Sensors do not provide exact readings of desired quantities, but introduce their own system dynamics and distortions as well.
- Furthermore, these devices are also always noise corrupted.

State Space Design

- Main idea: Motion of any finite dimensional dynamical system can be expressed as a set of first-order differential equations. Solution is visualized as a trajectory in space.
- Why is it useful?
  - We know a lot about the solution, stability of first-order ODEs.
  - This model format is particularly useful with regard to numerical computations.

- *Example*

  Consider the following spring-mass-damper system in Figure 1

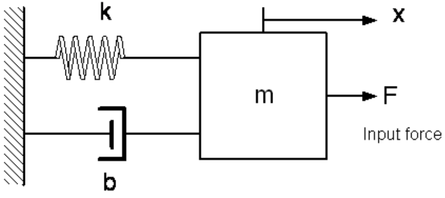


Figure 1: Mass-Spring-Damper System

The equation of motion (EOM) can be obtained from Newton's law of motion. Show that the EOM of Figure 1 is given by

$$m\ddot{x} = F - b\dot{x} - kx$$

Objective: To write the EOM as a set of first-order differential equations.

Step 1: Assign a "state variable" label "x_i" to each variable which has its derivatives represented in the EOM. Hence, for the above example, we have

state variable		variable in EOM
x_1	$:=$	x
x_2	$:=$	$\dot{x}$

We note that the states are not unique; however they are usually selected to have physical meaning (often directly describing the distribution of internal energy in a system). The common choice of state variable includes: position (potential energy), velocity (kinetic energy), capacitor voltage (electric energy) and inductor current (magnetic energy).

Step 2: Rewrite the EOM as a set of first-order d.e.'s in terms of the state variables

$$\begin{aligned} \dot{x_1} &= x_2 \\ \dot{x_2} &= -\frac{k}{m}x_1 - \frac{b}{m}x_2 + \frac{F}{m} \end{aligned}$$

And assuming that we are interested in the mass position x, and that we are given the initial condition on the state (i.e. at $t = t_0$, $x = x_0$), we have

$$y = x_1$$

Step 3: Now rewrite the above first-order d.e.'s into matrix form as

$$\begin{aligned} \dot{x} &= Fx + Gu &\longrightarrow& \quad \text{state equation} \\ y &= Hx + Ju &\longrightarrow& \quad \text{output equation} \\ x(t_0) &= x_0 &\longrightarrow& \quad \text{initial condition} \end{aligned} \tag{1}$$

where

$$x := \begin{bmatrix} \\ \\ \end{bmatrix}$$ state vector

$$u :=$$ input or control

$$F := \begin{bmatrix} \\ \\ \end{bmatrix}$$ dynamic matrix

$$G := \begin{bmatrix} \\ \\ \end{bmatrix}$$ input matrix

$$H := \begin{bmatrix} \\ \end{bmatrix}$$ output or sensor matrix

$$J := \begin{bmatrix} \\ \end{bmatrix}$$ feedthrough term

- In general, the state space representation of Equation 1 has the following dimensions
  - x is an $n \times 1$ state vector ($x \in \mathcal{R}^n$)
  - u is an $n_p \times 1$ state vector ($u \in \mathcal{R}^p$)
  - y is an $n_q \times 1$ state vector ($y \in \mathcal{R}^q$)
  - $F \in \mathcal{R}^{n \times n}$, $G \in \mathcal{R}^{n \times p}$, $H \in \mathcal{R}^{q \times n}$ and $J \in \mathcal{R}^{q \times p}$

  In this course, we are mainly interested in single input single output (SISO) systems. Hence, unless otherwise stated, $p = q = 1$.

Example

Electrical Models

- Basic equations of electric circuits are the Kirchhoff's laws:
  - Kirchhoff's current law (KCL)
  - Kirchhoff's voltage law (KVL)
- Example:
  Consider the electrical network in Figure 2. Using KCL, write down the ODE to describe the system. By choosing the inductor current, $i(t)$, and output voltage, $v_o(t)$, as the state variables, obtain the state-space model.

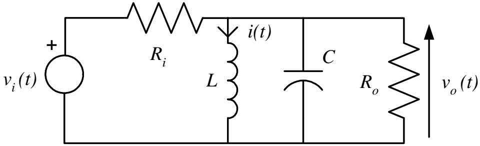


Figure 2: Electrical network

- Let $x_1(t) = i(t)$ and $x_2(t) = v_o(t)$, using KCL, we have

$$\begin{aligned} x_2(t) = v_o(t) &= L\frac{di(t)}{dt} \\ &= L\dot{x}_1(t) \\ \dot{x}_1(t) &= \frac{1}{L}x_2(t) \end{aligned}$$

and

$$\begin{aligned} \frac{v_i(t) - v_o(t)}{R_i} &= i(t) + \frac{v_o(t)}{R_o} + C\frac{dv_o(t)}{dt} \\ \Rightarrow \dot{x}_2(t) &= -\frac{1}{C}x_1(t) - \left(\frac{1}{R_oC} + \frac{1}{R_iC}\right)x_2(t) + \frac{1}{R_iC}v_i(t) \end{aligned}$$

- Let $v_i(t) = u(t)$, we have in state-space format:

$$
\begin{aligned}
\begin{bmatrix} \dot{x_1}(t) \\ \dot{x_2}(t) \end{bmatrix} &= \begin{bmatrix} 0 & \frac{1}{L} \\ -\frac{1}{C} & -\left(\frac{1}{R_o C} + \frac{1}{R_i C}\right) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{R_i C} \end{bmatrix} u(t) \\
y(t) &= \begin{bmatrix} 0 & 1 \end{bmatrix} x(t) + \begin{bmatrix} 0 \end{bmatrix} u(t)
\end{aligned}
$$

Motivations

- Sometimes a physical variable/quantity that we wish to measure is inaccessible or not cost-effective to measure directly.
  - commercial sensor might not be available
  - environment too severe to allow a sensor to survive
- In state-variable controller design (see Figure 3, the control engineer designs a dynamic compensation by working directly with the state-variable description of the system. This approach assumes that all state variables are measureable (not a practical assumption).
- Rarely able to directly measure all the state variables $\Longrightarrow$ use of an estimator/observer to reconstruct all the state variables of a system from a few measurements.

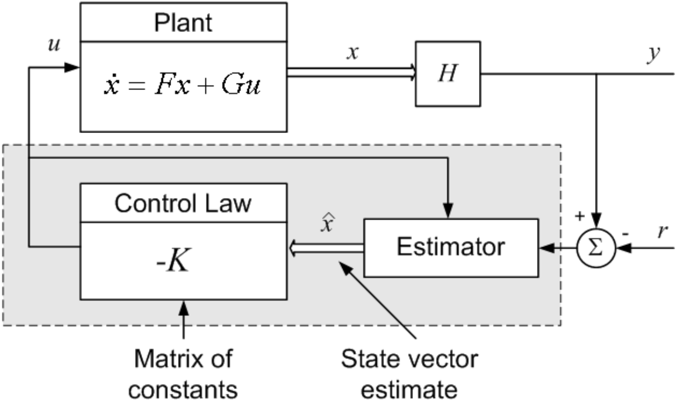


Figure 3: Schematic diagram of state-space design elements.

Estimator Design

- In its simplest form, the estimator/observer is nothing but a mathematical model of the phsyical system in which the inaccessible variables are found, $\implies$ the model of the physical system in important.

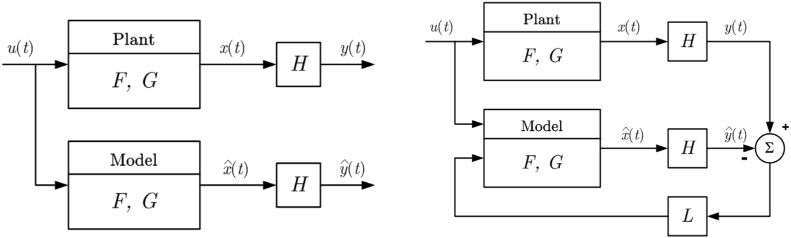


Figure 4: Types of estimator/observer: open-loop (left), closed-loop (right)

- It turns out that it is alright to use an estimate $\hat{x}$ of the state x as long as
$$\hat{x} - x \to 0 \quad \text{exponentially}$$
so that the new control law is given by
$$u = -K\hat{x}$$

- *Open-loop Estimator*
For the following system,
$$\begin{aligned} \dot{x} &= Fx + Gu \\ y &= Hx \end{aligned} \tag{2}$$
Consider the estimator model of the plant dynamics,
$$\dot{\hat{x}} = F\hat{x} + Gu; \qquad \hat{x}(0) \tag{3}$$
Since we know F, G and u in equation 3, this estimator will work if we can obtain the correct $x(0)$ and set $\hat{x}(0)$ to it. Such an estimator is "open-loop" since no information on the prediction error is being feedback to obtain a better $\hat{x}(k)$.

The estimation error is given by

$$\tilde{x} = x - \hat{x} \tag{4}$$

Substituting Equations 2 and 3 into 4, the dynamics of the resulting system are described by the estimator-error equation

$$\dot{\tilde{x}} = F\tilde{x}, \qquad \tilde{x}(0) = x(0) - \hat{x}(0) \tag{5}$$

Thus, if the initial value of $\hat{x}$ is off, the dynamics of the estimate error are those of the uncompensated plant, F. The error converges to zero for stable system (F stable), but we basically have no ability to influence the rate at which the state estimate converges to the true state.

- *Example*: Consider the simple case where $n = 1$, $x(0) = 1$ and $\hat{x}(0) = 0$. We then have

$$\dot{\tilde{x}} = f\tilde{x}; \quad \tilde{x}(0) = 1$$

and

$$\tilde{x}(t) = e^{ft}\tilde{x}(0)$$

Thus

$$\tilde{x}(t) \to 0 \quad \text{only if} \quad f < 0$$
$$\tilde{x}(t) \quad \text{grows exponentially if} \quad f > 0$$

Therefore, open-loop estimator

- cannot be used for systems with any open-loop unstable pole;
- even if all open-loop poles are stable, estimation error $\tilde{x}$ may not decrease fast enough if there are slow poles in the system.

A better estimator would thus be one that make use of the error information.

- *Closed-loop Estimator*

  Figure 4 show such a scheme where the difference between the measured output and the estimated output is being feedback and constantly correcting the model with this error signal.

  The dynamics of the closed-loop estimator is given by

$$\dot{\hat{x}} = F\hat{x} + Gu + L(y - H\hat{x}) \tag{6}$$

where

$$L = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_n \end{bmatrix}$$

is an estimator gain chosen to achieve satisfactory dynamics of the error $\tilde{x}$.

- From Equations 2 and 6, the error dynamics is now given by

$$\dot{\tilde{x}} = (F - LH)\tilde{x} \tag{7}$$

and the poles of the error system are at

$$\det\left[sI - (F - LH)\right] = 0 \tag{8}$$

Fast poles would correspond to faster decay of the error.

- *Example*: Continuing with our first-order system example, assume that $f = 1$, and $y = 2x$
  Using the estimator $\dot{\hat{x}} = \hat{x} + u + l(y - 2\hat{x})$, the error dynamics is

  $$\dot{\tilde{x}} = (1 - l \cdot 2)\tilde{x}$$

  Choosing $l = 2$, we have

  $$\begin{aligned} \dot{\tilde{x}} &= -3\tilde{x} \\ \tilde{x}(t) &= e^{-3t}\tilde{x}(0) \end{aligned}$$

  (Obviously, this example is rather constrained; but the point should be clear.)

Therefore, L should be chosen so that the eigenvalues of $F - LH$ are stable, and sufficiently fast.
We note here that the plant is a physical system such as a chemical process or servomechanism, whereas the estimator is usually an electronic unit computing the estimated state according to the estimator equation.

Example 1

- Simple spring, mass and damper system.
  - Consider the system shown in Figure 5, assume measurement of force input and mass displacement are available but mass velocity is not possible.

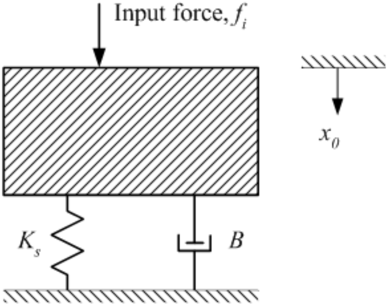


Figure 5: Spring-mass-damper system.

- Newton's law will give us a second-order differential equation which can be expressed as two first-order differential equations. The observer gains for the two states, x_0 and v_0 are 30 and 10 respectively (manually tuned to provide good response for the estimator). Figure 6 and 7 shows the displacement and velocity of the two states.
- Note that the estimator is turned on from time equal to 1 second. Our simulation shows both the measured displacement and velocity, however, only the displacement information is fed into the estimator. The simulation is conducted with a 20% error in the observer/estimator model.

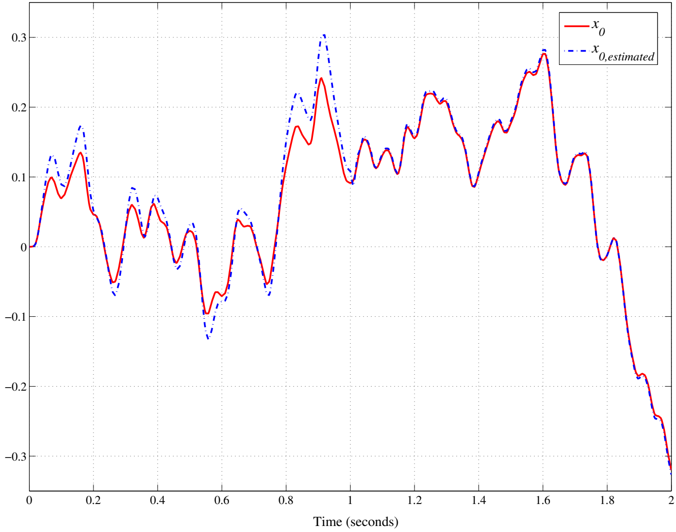


Figure 6: Displacement measurement and estimation.

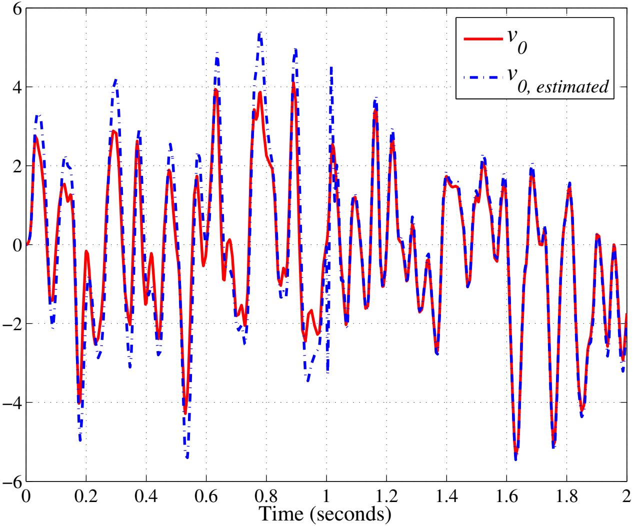


Figure 7: Velocity measurement and estimation.

Example 2

- Real-time control of wafer temperature during thermal processing in microlithography.
  - Reference: A. Tay et al., "Modeling and Real-Time Control of Multizone Thermal Processing System for Photoresist Processing", Industrial & Engineering Chemistry Research, 52, 4805, 2013.
  - Consider Figure 8 shows a thermal processing system used for photoresist processing in microlithography in the fabrication of integrated circuits. The temperature on the wafer is of interest and important to the final linewidth of the circuit. Variation in temperature will lead to variation in the linewidth.
  - Our objective is to control the temperature on the wafer in real-time. However, it is not possible to have sensors on the wafer. Non-contact sensors are also difficult to install because most of these bake systems has a cover. The only available temperature sensors are on the bake-plate.

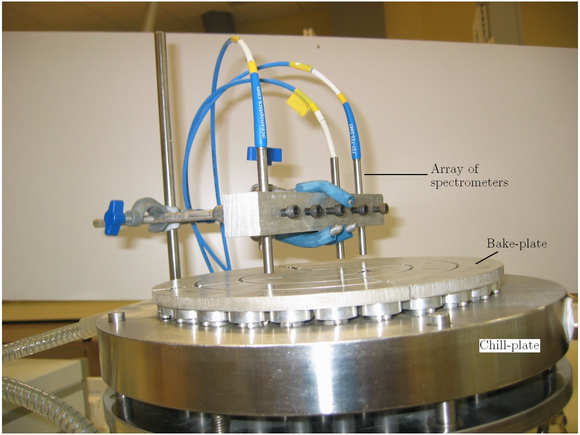


Figure 8: Integrated bake/chill system for thermal processing in lithography.

- In this work, a thermal model of the system is first constructed based on energy balance equation as follows:

$$
\begin{aligned}
C_{p1}\dot{\theta}_{p1} &= -\frac{\theta_{p1}-\theta_{p2}}{R_{p1}} - \frac{\theta_{p1}-\theta_{w1}}{R_{a1}} + q_1, \\
C_{pi}\dot{\theta}_{pi} &= \frac{\theta_{p(i-1)}-\theta_{pi}}{R_{p(i-1)}} - \frac{\theta_{pi}-\theta_{p(i+1)}}{R_{pi}} - \frac{\theta_{pi}-\theta_{wi}}{R_{ai}} + q_i, \qquad 2 \leq i \leq N-1, \\
C_{pN}\dot{\theta}_{pN} &= \frac{\theta_{p(N-1)}-\theta_{pN}}{R_{p(N-1)}} - \frac{\theta_{pN}}{R_{pN}} - \frac{\theta_{pN}-\theta_{wN}}{R_{aN}} + q_N, \\
C_{w1}\dot{\theta}_{w1} &= \frac{\theta_{p1}-\theta_{w1}}{R_{a1}} - \frac{\theta_{w1}-\theta_{w2}}{R_{w1}} - \frac{\theta_{w1}}{R_{wz1}}, \\
C_{wi}\dot{\theta}_{wi} &= \frac{\theta_{w(i-1)}-\theta_{wi}}{R_{w(i-1)}} + \frac{\theta_{pi}-\theta_{wi}}{R_{ai}} - \frac{\theta_{wi}-\theta_{w(i+1)}}{R_{wi}} - \frac{\theta_{wi}}{R_{wzi}}, \quad 2 \leq i \leq N-1, \\
C_{wN}\dot{\theta}_{wN} &= \frac{\theta_{w(N-1)}-\theta_{wN}}{R_{w(N-1)}} + \frac{\theta_{pN}-\theta_{wN}}{R_{aN}} - \frac{\theta_{wN}}{R_{wN}} - \frac{\theta_{wN}}{R_{wzN}}.
\end{aligned}
$$

where

$\theta_{pi} = T_{pi} - T_\infty$ is the ith plate element temperature above ambient;

$\theta_{wi} = T_{wi} - T_\infty$ is the ith wafer element temperature above ambient;

T_∞ is the ambient temperature;

C_{pi} is the thermal capacitance of ith plate element;

C_{wi} is the thermal capacitance of ith wafer element;

R_{pi} is the thermal conduction resistance between the ith and $(i+1)$th plate element;

R_{wi} is the thermal conduction resistance between the ith and $(i+1)$th wafer element;

R_{wzi} is the thermal convection loss of the ith wafer element;

R_{ai} thermal conduction resistance between the ith plate and ith wafer element; and

q_i is the power into the ith plate element.

- The various thermal resistances and capacitances are given by

$$\begin{aligned}
R_{pi} &= \frac{\ln\left(\frac{i+1/2}{i-1/2}\right)}{2\pi k_a t_p} \quad (K/W), && 1 \leq i \leq N-1, \\
R_{pN} &= \frac{1}{h(\pi D t_p)} \quad (K/W), \\
R_{wi} &= \frac{\ln\left(\frac{i+1/2}{i-1/2}\right)}{2\pi k_w t_w} \quad (K/W), && 1 \leq i \leq N-1, \\
R_{wN} &= \frac{1}{h(\pi D t_w)} \quad (K/W), \\
R_{wzi} &= \frac{1}{h A_{zi}} \quad (K/W), \\
R_{ai} &= \frac{t_{ai}}{k_a A_{zi}} \quad (K/W), \\
C_{pi} &= \rho_p c_p \left(t_p A_{zi}\right) \quad (J/K), && 1 \leq i \leq N, \\
C_{wi} &= \rho_w c_w \left(t_w A_{zi}\right) \quad (J/K), && 1 \leq i \leq N, \\
A_{zi} &= \pi \Delta_r^2 \left[i^2 - (i-1)^2\right] \quad (m^2), && 1 \leq i \leq N.
\end{aligned}$$

where A_{zi} is the cross-sectional area of element i normal to the axial heat flow. t_p and t_w are the bake-plate thickness and wafer thickness respectively and t_{ai} is the air-gap between the ith wafer and bake-plate elements. k_a and k_w are the thermal conductivity of air and wafer respectively. h is the convective heat transfer coefficient. ρ_p and ρ_w are the density of the bake-plate and wafer respectively. c_p and c_w are the specific heat capacity of the bake-plate and wafer respectively. The width of each element is given by $\Delta_r = D/(2N)$.

The above energy balance equations can be rearranged into state-space format.

$$\begin{aligned}\dot{\boldsymbol{\theta}} = \begin{bmatrix} \dot{\theta}_p \\ \dot{\theta}_w \end{bmatrix} &= \begin{bmatrix} \mathbf{F_{pp}} & \mathbf{F_{pw}} \\ \mathbf{F_{wp}} & \mathbf{F_{ww}} \end{bmatrix} \begin{bmatrix} \theta_p \\ \theta_w \end{bmatrix} + \begin{bmatrix} \mathbf{G_{pp}} \\ \mathbf{0_N} \end{bmatrix} \mathbf{q} \qquad (9) \\ &= \mathbf{F}\theta + \mathbf{G}\mathbf{q},\end{aligned}$$

where θ_p and θ_w are vectors containing the n-plate and wafer temperatures respectively. The state matrices $\mathbf{F}$ and $\mathbf{G}$ can be found in the reference paper.

- Once we have the model, an estimator can be used to estimate and control the wafer temperature as shown in Figures 9 and 10. In this work, the state-space model has to be re-identified during each wafer runs, the model is affected due to wafer warpages.
- We note that for the identification work, only the plate temperatures and the heater input signal is required. The wafer temperature remains unknown. Subspace identification techniques can be employed.

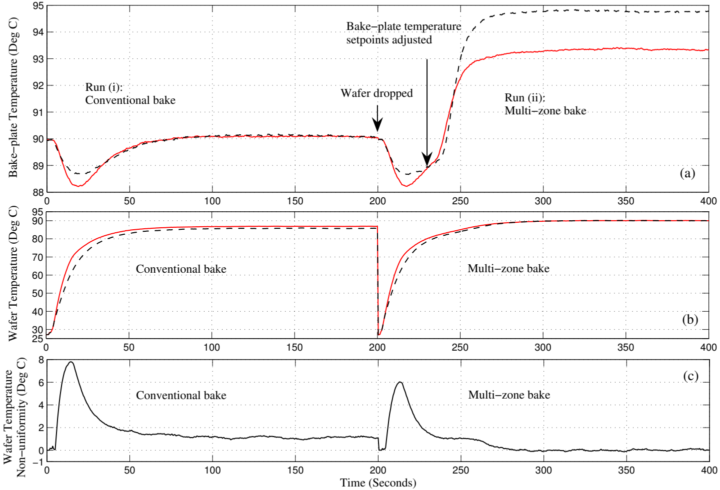


Figure 9: Temperature profile of bake-plate and wafer when a wafer with center-to-edge warpage of 110 μm is dropped on bake-plate with proximity pin height of 220 μm. The bakeplate temperatures, wafer temperatures and wafer temperature non-uniformity during the baking process are shown in subplots (a), (b) and (c) respectively.

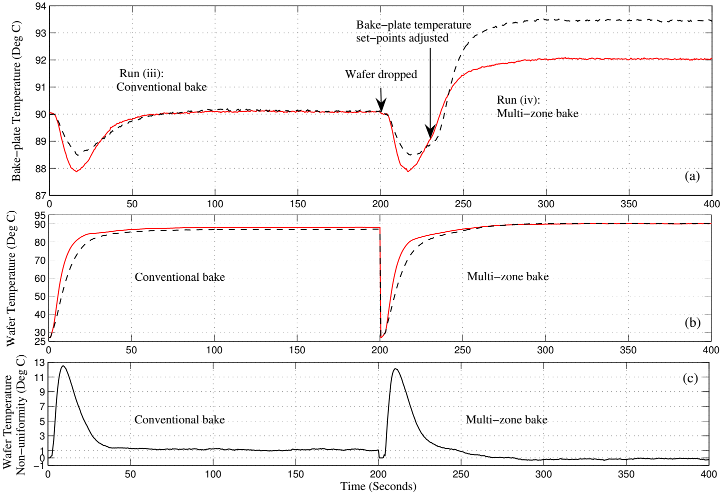


Figure 10: Temperature profile of bake-plate and wafer when a wafer with center-to-edge warpage of 110 μm is dropped on bake-plate with proximity pin height of 165 μm. The bakeplate temperatures, wafer temperatures and wafer temperature non-uniformity during the baking process are shown in subplots (a), (b) and (c) respectively.

Reduced-order Estimator

- In many systems, we can usually measure quite a few states, hence it make sense if we just re-construct the unmeasurable states.
- Consider the following system

$$
\begin{aligned}
\dot{x} &= Fx + Gu \\
y_m &= \begin{bmatrix} H_a & \vdots & 0 \end{bmatrix} \begin{bmatrix} x_a \\ \cdots \\ x_b \end{bmatrix}
\end{aligned}
$$

  where x_a and y_m is an n_a-dimension measurement vector, H_a is nonsingular $n_a \times n_a$ matrix.
- Note that we are allowing y_m to be n_a-dimensional. This is reasonable because several measurements may be avaliable.

- For the system

$$\begin{bmatrix} \dot{x}_a \\ \dot{x}_b \end{bmatrix} = \begin{bmatrix} F_{aa} & F_{ab} \\ F_{ba} & F_{bb} \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix} + \begin{bmatrix} G_a \\ G_b \end{bmatrix} u$$

$$y_m = \begin{bmatrix} H_a & 0 \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix}$$

  - x_a is recovered directly as $x_a = H_a^{-1} y_m$
  - x_b is estimated by $\hat{x}_b = L y_m + z$ with $\dot{z} = \bar{F} z + \bar{G} y_m + \bar{H} u$

where

$$\begin{aligned} \bar{F} &= F_{bb} - L H_a F_{ab} \\ \bar{G} &= \left(F_{ba} - L H_a F_{aa} + \bar{F} L H_a \right) H_a^{-1} \\ \bar{H} &= G_b - L H_a G_a \end{aligned}$$

and L is chosen to ensure that eigenvalues of $\bar{F}$ are all in the l.h.p.

Example 3

- Consider the undamped oscillator

$$\begin{aligned}\dot{x} &= \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} x\end{aligned}$$

To design a reduced-order estimator to place the error pole at $-10\omega_0$. The partitioned matrices are

$$\begin{bmatrix} F_{aa} & F_{ab} \\ F_{ba} & F_{bb} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{bmatrix}, \qquad \begin{bmatrix} G_a \\ G_b \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The c.e. of the system is

$$\begin{aligned}\det\left[sI - (F_{bb} - LH_aF_{ab})\right] &= 0 \\ \det\left[s - (0 - L \cdot 1 \cdot 1)\right] &= 0 \\ s + L &= 0\end{aligned}$$

- The desired c.e. is $\alpha_e(s) = s + 10\omega_0 = 0$. Comparing, we have $L = 10\omega_0$. In Matlab, this can be generated using

$$L = \text{acker}(F_{bb}^T, F_{ab}^T H_a^T, \rho_0)$$

- We have

$$\begin{aligned}
\bar{F} &= F_{bb} - LH_aF_{ab} \\
&= 0 - 10\omega_0 \cdot 1 \cdot 1 \\
\bar{G} &= \left(F_{ba} - LH_aF_{aa} + \bar{F}LH_a\right) H_a^{-1} \\
&= (-\omega_0^2 - 10\omega_0 \cdot 0 + (-10\omega_0) \cdot 10\omega_0 \cdot 1) \cdot 1^{-1} \\
&= -101\omega_0^2 \\
\bar{H} &= G_b - LH_aG_a \\
&= 1 - 10\omega_0 \cdot 1 \cdot 0 = 1
\end{aligned}$$

Thus x_b is estimated by

$$\begin{aligned}
\hat{x}_b &= 10\omega_0 y_m + z \\
\dot{z} &= -10\omega_0 z - 101\omega_0^2 y_m + u
\end{aligned}$$

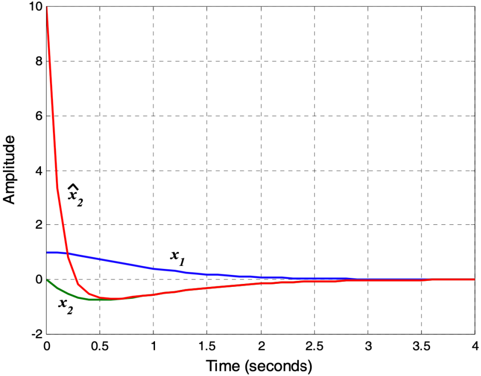
- Figure below shows the initial condition response of the reduced-order estimator to $\omega_0 = 1$, $x_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$ and $z_0 = 0$.

The Matlab code are as follows:

```
% The open-loop system
f = [0 1;-1 0]; g = [0;1]; h = [1 0];
K = [3 4];

% The combined system
A = [f-g*K [0;0]; -101*h-K -10]; B = zeros(3,1);
C = [h 0]; D = 0;
syscl = ss(A,B,C,D);

t = 0:.1:4;
[y,t,x] = initial(syscl,[1;0;0],t);
plot(t,x(:,1),'r',t,x(:,2),'b',t,10*x(:,1)+x(:,3),'g');
xlabel('Time (seconds)');
ylabel('Amplitude');
```



How to choose L?

Design rules of thumb

- Choose estimator poles 2 to 6 times faster than closed-loop poles from control law design. This will ensure that the controller poles dominate the total response.
  - the faster the desired response, in general the larger the corresponding feedback signal
  - Control law: $u = -K\hat{x}$

    fast $\alpha_c(s) \Rightarrow$ large u

    actuator may saturate
  - Estimator: feedback signal to estimator is $L(y - H\hat{x})$

    fast $\alpha_e(s) \Rightarrow$ large $L(y - H\hat{x})$ signal

    this is alright as this signal is only a voltage in analog implementation or digital word in discrete implementation – no difficulty

    only limitation is that higher estimator bandwidth implies more sensor noise to pass on to the control actuator – undesirable
  - large sensor noise, choose slower estimator poles.
- An estimator/observer is basically a mechanism to estimate the unmeasurable states from available output measurements – a special form of virtual sensor.